A Study of Automated Web Site Evaluation Tools

Melody Y. Ivory
The Information School
University of Washington
myivory@u.washington.edu

Aline Chevalier
Department of Cognitive Psychology
University of Provence
alinech@up.univ-aix.fr

Technical Report UW-CSE-02-10-01
October 8, 2002

Abstract

Web site usability and accessibility continues to be a pressing problem. Hence, there are over 30 automated evaluation tools to help designers to improve their sites. Unfortunately, there is little evidence about whether these tools actually improve sites from both the designer's and the user's perspectives. To gain insight, we conducted what we believe to be the first study of automated web site evaluation tools. We describe current practices for creating usable and accessible sites. We present findings from experienced designers' usage of three automated evaluation tools to improve sites, and we present findings from users' usage of the original and modified sites. We also discuss study implications.

Keywords: World Wide Web, Empirical Studies, Automated Usability Evaluation, Web Site Design, Universal Access

Introduction

The WWW has become the predominate means for communicating and presenting information on a broad scale. Unfortunately, despite the abundance of design recommendations and guidelines for building effective sites, web site usability and accessibility continues to be a pressing problem. Use of automated tools to evaluate and ideally improve some aspects of web sites is one way to address this problem. Currently, there are over 30 automated tools, such as WatchFire Bobby,[1] UsableNet LIFT,[2] and the W3C HTML Validator,[3] for evaluating web site usability, accessibility, coding, etc. [7].

Despite their potential benefits, there is little evidence about the efficacy of these tools, specifically whether they: (1) from a designer's perspective, result in sites that are better than sites produced without tools; and (2) from a user's perspective, result in sites that are more usable and accessible than sites produced without tools. We are not aware of any other studies that assess automated evaluation tools along these dimensions. However, Brazier and Jennings [3] discuss some of the limitations of the Bobby tool (e.g., the lack of guidance on the structure of the entire web site), and Brajnik [2] claims that most tools assess only a sparse set of usability features. Furthermore, many tools analyze the HTML code and rarely the design itself [7].

This paper presents an empirical study of three automated web site evaluation tools. We discuss background information and related work. We then summarize responses to a survey on the work practices of web professionals, including their use of automated evaluation tools. We report findings from a study wherein experienced web designers used the WatchFire Bobby, W3C HTML Validator, and UsableNet LIFT tools to modify five sites. We then report findings from a study wherein users with and without visual, physical, and learning impairments completed information-seeking tasks on the original and modified sites. We discuss study implications for both researchers and web professionals.

[1] Information about the WatchFire Bobby tool is available at http://bobby.watchfire.com/bobby/html/en/index.jsp.

[2] Information about the UsableNet LIFT tools is available at http://www.usablenet.com/products_services/products_services.html.

[3] Information about the W3C HTML Validator tool is available at http://validator.w3.org/.

Background and Related Work

An automated evaluation tool is software that automates the collection of interface usage data (automated capture) or the identification (automated analysis) and the resolution (automated critique) of potential problems [7]. Ivory and Hearst [7] surveyed 58 methods for automated evaluation of web interfaces and found that most approaches entailed analyzing server and other log file data (log file analysis) or determining whether an interface conforms to a set of usability, accessibility, or other guidelines (guideline review). We examine three guideline review tools that support automated critique.

As far as we know, there is little evidence about the efficacy of the existing automated web site evaluation tools, though there have been studies of a few. For example, Blackmon and colleagues [1] showed that their methodology for simulating information-seeking behavior could predict potential problems with confusing headings and link text. Researchers on the WebTango project [8] showed that participants preferred web pages that were modified to conform to their statistical models of highly rated web designs over the original ones and that participants rated sites modified based on their models higher than the original ones; differences were significant in both cases. Lynch, Palmiter, and Tilt [9] provided preliminary evidence that the model of navigation time embedded in WebCriteria's SiteProfile[4] corresponded roughly to observed behavior. UsableNet's LIFT - Nielsen Norman Group Edition[5] assesses a site's conformance to guidelines developed from studies involving users with and without visual and mobility impairments [5].

Deaton[6] asked eight web designers to rank a list of 50 guidelines (e.g., avoid excessive scrolling and be consistent in the use of fonts) on a seven-point scale (always do to never do). Participants' rankings revealed that practices involving consistency were most frequently used, but practices involving accessibility were seldom used. Chevalier and Ivory [4] conducted a study wherein fourteen professional and novice web designers produced initial designs for a car dealer's site in conditions with and without eleven design constraints (e.g., the web site must be short: 10–15 pages maximum). The study revealed that all designers experience difficulties with implementing design constraints and consequently need additional support, perhaps via a guideline review tool.

Web Professionals' Work Practices

We surveyed 169 web professionals who had a role in developing sites for use by broad user communities. We recruited participants by sending email announcements to mailing lists (CHI-Web, ASIS IA, ASIS HCI, and many others) and by posting announcements on web sites, such as SpiderMetrix.com. We asked participants about their work practices related to creating sites that are universally accessible (i.e., allow "access to information by individuals with different abilities, requirements, and preferences, in a variety of contexts of use" [10])[7] and usable (i.e., the extent to which users can use a site to achieve specified goals effectively and efficiently while promoting feelings of satisfaction in a given context of use; adapted from [6]). We asked survey respondents about their usage and assessment of twenty-five automated evaluation tools and asked them to explore and rate the universal accessibility of ten sites on a five-point scale from very high to very low.

Almost an equal number of males and females responded to the survey. They represented a broad range of practitioners from web designers (12.4%) to information architects/designers (23.1%) and usability analysts/human factors (14.8%). They had varying design expertise (beginner: 30.8%, intermediate: 50.9%, and expert: 18.3%) with 57.4% of them having worked in the field for more than three years and 45.6% of them having designed more than ten sites. Participants worked in many environments, including corporations, academic institutions, and independently; they worked on education, e-commerce, intranet, health, and other types of sites.

[4] This tool is no longer available for use.

[5] Information about the UsableNet LIFT - Nielsen Norman Group Edition is available at http://www.usablenet.com/products_services/lfdnng/lfdnng.html.

[6] Personal communication, Mary Deaton, *Do as We Say or as We Do? Discovering Standard Practices Amongst Web Practitioners*, 2002.

[7] We think that a universally accessible web site is one that is both usable and accessible; however, this view may not be shared across web designers and users.

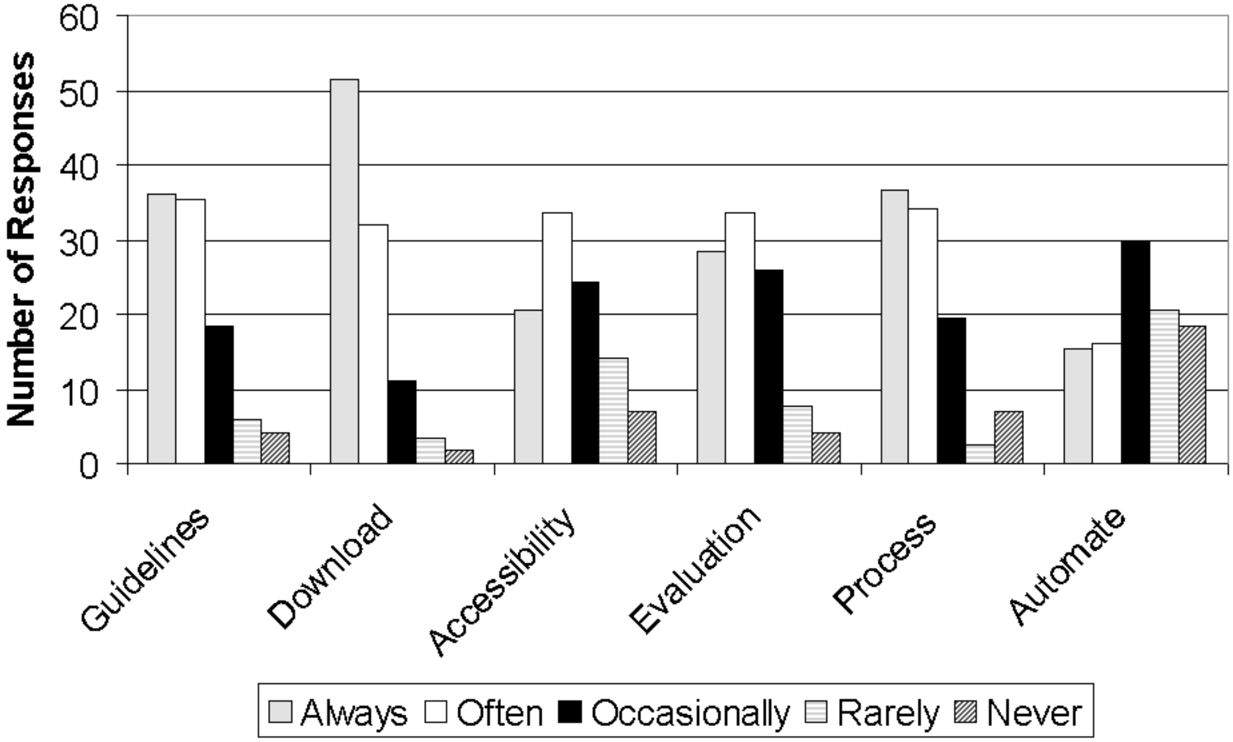

Figure 1: Self-reported practices for creating universally accessible sites: using design guidelines/style guides (Guidelines), optimizing sites for fast download (Download), optimizing sites to improve access by people with disabilities (Accessibility), testing with users (Evaluation), using a user-centered design/development process (Process), and using automated evaluation tools (Automate).

Use of Automated Evaluation Tools

Figure 1 summarizes some of the practices that web professionals follow to create universally accessible sites, including using design guidelines, testing with users, using a user-centered design process, and using automated evaluation tools. Over half of the respondents reported that they always or often follow all practices, except for using automated evaluation tools. We asked the 137 respondents who reported using automated evaluation tools about the design stages in which they use them; 54.4% reported using them throughout the design process as compared to using them early in the design process (5.9%) or after finishing the site (20.7%).

Figure 2 shows that the respondents had mixed opinions about the automated evaluation tools they had used. At least half of them thought that the tools were helpful in creating universally accessible sites (Help) and in learning about effective design practices (Teach). However, the majority of them did not think that using the tools produced better sites than those produced without using tools (Better). They did not think that the tools had adequate functionality or support (Adequate) or that they were easy to use (Easy). Participants' responses suggest that it is helpful for them to have another way to check their designs, after making in particular, but the tools are too difficult and too limited to use.

Most respondents reported using the WatchFire Bobby[8] and the W3C HTML Validator guideline review tools (47% and 43% of respondents, respectively). Other frequently used tools include: the Dreamweaver 508 Accessibility Suite (16%),[9] AnyBrowser (11%),[10] WAVE (9%),[11] and LIFT for Dreamweaver (9%).

We asked the 137 respondents, who had reported using automated evaluation tools, to identify both the best and the worst tools they had used. Respondents identified Bobby most often as the best tool (27%) and

[8] Researchers at the Center for Applied Special Technology originally developed Bobby; WatchFire acquired it in July 2002.

[9] Information about the Macromedia Dreamweaver 508 Accessibility Suite is available at http://www.macromedia.com/macromedia/accessibility/.

[10] Information about the AnyBrowser tool is available at http://www.anybrowser.com/siteviewer.html.

[11] Information about the WAVE tool is available at http://www.temple.edu/inst_disabilities/piat/wave/.

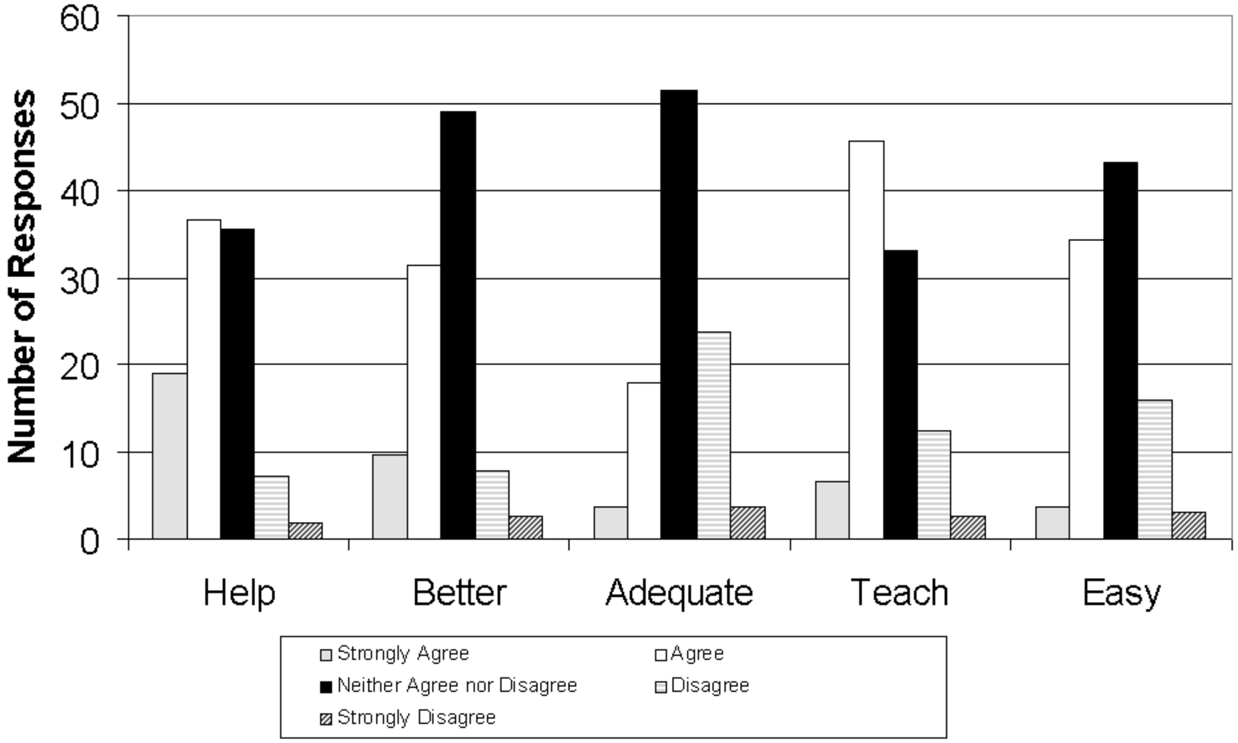

Figure 2: Respondents' responses to statements about the automated evaluation tools they had used: helpful in creating universally accessible web sites (Help), create better sites with tools (Better), adequate functionality/support (Adequate), help users to learn about effective design practices (Teach), and easy to use (Easy).

as the worst tool (9%). Many respondents stated that they selected Bobby as the best tool, because it was the only one they knew of or had used. Many respondents also thought that it provided more comprehensive information, liked the multiple guidelines it evaluates, and thought that it was relatively easy to use and pass; however, they also mentioned that it was not easy enough to use and that it did not provide enough information. Many respondents stated that they selected Bobby as the worst tool, because it was not easy to use, they disagreed with its guidelines, and it did not provide information on how to fix problems.

Assessment of Web Sites

To identify web sites for a subsequent study, we asked survey respondents to explore, rate, and comment on the universal accessibility of ten web sites. From these ten sites, we selected five—www.section508.gov, www.gov-benefits.gov, spam.abuse.net, www.irs.gov, and www.-usatoday.com—for a subsequent study because of: respondents' ratings, respondents' identification of several usability or accessibility problems on them, the sites' potential interest to a broad user community, and the sites' implementation predominantly in static HTML. Three of the sites were government sites, which meant that they had to adhere to the Section 508 requirements.

Web Designers' Modification of Web Sites

We conducted a study with nine experienced designers wherein each designer modified subsections of the five sites; modifications were intended to improve universal accessibility. Designers completed modifications with or without the assistance of one or more automated evaluation tool (WatchFire Bobby, UsableNet LIFT, or W3C HTML Validator). Our objective was to examine designers' usage of these frequently used tools and to assess whether using the tools was beneficial to them.

Study Design

We hypothesized that using an automated evaluation tool to guide site modifications would enable designers to identify more site problems, to correct more of these problems, and to consequently better improve the usability and accessibility of the site subsections, at least more so than not using automated evaluation tools. To test this hypothesis, we asked designers to participate in a 5x5 between-subjects experiment wherein each designer modified all five web site subsections; they modified each site in one of five evaluation conditions. Designers completed the first site modification without the assistance of an automated evaluation tool (manual). Designers completed the second through fourth site modifications using one automated evaluation tool (Bobby, Validator, or LIFT). Designers completed the fifth and final site modification using all three tools (combo). We counterbalanced both the order of site modifications and the usage of individual tools.

Designers completed the following tasks for each site and evaluation condition:

1. Explore the site to become familiar with the subsection. We provided designers with a usage scenario (Table 1) and disabled links within the subsection to other site areas or to external sites.

2. Evaluate the universal accessibility of the site subsection. Designers used a five-point scale (very high to very low) for these ratings. We did not specify how designers were to make these judgments; in some cases, designers inspected both the rendered pages and the HTML code.

3. Modify the site subsection to improve its universal accessibility. In conditions without an automated evaluation tool, designers enumerated a list of problems and used their own expertise to guide modifications. Otherwise, they could only address problems identified by the tool. Designers made modifications using either the Microsoft FrontPage or the Macromedia Dreamweaver authoring environments on PCs running Microsoft XP.

4. Evaluate the universal accessibility of the modified site subsection (same rating scale used in step 2).

5. Evaluate the automated evaluation tool (conditions with one tool).

There were at least eight site modifications in each of the five conditions—manual, bobby, validator, lift, or combo. Study sessions were three hours and due to time constraints, we asked designers to spend twenty minutes modifying the first four sites and to spend forty minutes modifying the final site (combo condition). Designers completed a debriefing questionnaire about their experiences with using the tools. We asked them think aloud throughout the study session, and we videotaped sessions and captured screen activity.

Table 1 summarizes the site subsections that designers rated and modified. Site contents were of general interest to a broad user community, with three of the sites—1, 2, and 4—being government sites. Site designs ranged from a simplistic design with few graphics (site 1) to a text- and graphics-intensive design (site 5).

Study Tools

The WatchFire Bobby, UsableNet LIFT, and W3C HTML Validator tools represent the state-of-the-art in automated web site evaluation. They were among the most frequently used tools, offered different functionality, and enabled subsequent evaluation of site changes.

WatchFire Bobby: Determines if web pages conform to the W3C Content Accessibility or Section 508 guidelines and provides guidance for correcting problems. Example guidelines include: provide alternative text for all images and use relative sizing and positioning (% values) rather than absolute (pixels). Designers used the desktop software (v4.0), which enables them to crawl and evaluate all pages within a site. It presents an initial summary of the number of errors identified by each type: priority one (P1), two (P2), and three (P3). There is also a detailed report, which contains error descriptions, potential solutions, and manual checks to complete.

W3C HTML Validator: Determines if web pages conform to W3C HTML coding standards and provides guidance for correcting violations. Example guidelines include: there is no attribute "LEFTMARGIN" for this element (BODY tag) and element "TR" not allowed here. Designers used a copy of the web-based tool (June 22, 2001 version) that we installed on a local web server. The tool provides a list of violations, wherein each violation has HTML code information and an error description. Designers have to specify each URL to validate, and the tool does not support site crawling. If the tool is unable to identify a document

Site	Id	P	Subsection	Scenario
Section 508	1	8	About 508	Making a company's site EEOC compliant
Gov Benefits	2	8	GovBenefits Program List	Finding funding sources for a clinical geriatrics professor
Spam Abuse	3	6	Help for Users	Complaining about rude spam emails
IRS	4	7	Accessibility	Finding a Braille version of the 2001 Education Credits form
USA Today	5	6	FAQ	Getting the crossword puzzle to work in AOL

Table 1: Description of the five study sites, including an identifier (Id), the number of pages in the subsection (P), the name of subsection, and the usage scenario. Three pages on the Spam Abuse site were ASCII text rather than HTML pages.

type, then it reports a fatal error and does not attempt to validate the page. It reported a fatal error for two pages on site 2 and for all pages on sites 3, 4, and 5.

UsableNet LIFT: Determines if web pages conform to accessibility and usability guidelines (e.g., W3C Content Accessibility and Section 508), provides guidance for correcting problems, and in some cases, provides support for repairing the HTML code for images, tables, and forms. Unlike the other tools, designers used versions of the tool embedded within FrontPage or Dream-weaver. The developers claim that the Dreamweaver version, LIFT Nielsen Norman Group Edition (LIFT-NNG), contains the same core guidelines as the FrontPage version (LIFT-FP) as well as additional guidelines derived from usability studies involving participants with and without visual and mobility impairments [5]. Example guidelines include: use 'skip links' so that users can skip links or navigational elements (both versions) and avoid too many outgoing links (LIFT-NNG).

Study Participants

Nine designers, mostly recruited from the survey respondents, participated in the study. All designers had worked in the field for at least two years, with 66.7% of them having worked more than three years (all had designed at least four to ten sites) and 33.3% having worked two to three years (most of them had designed one to three sites). At least 66.7% of designers had never used any of the study sites. At least 75% of them had not used Bobby and only 22% had not used the Validator. Only one designer had used the LIFT tool. Designers had to be comfortable with modifying web pages within Dreamweaver or FrontPage.

Study Data

Study data include the tool evaluations, the site evaluations, and the responses to the debriefing questionnaire. In addition, we ran all tools (with default options) on the original and the modified pages; this enabled us to determine the number and the nature of problems both before and after modifications. We analyzed the problems that designers reported in the manual condition similarly. We used the ComponentSoftware HTML-Diff tool[12] to assist us with quantifying and describing modifications. We counted element-level changes; for instance, we counted adding space before a text link, changing its font size, and wrapping it across two lines as only one change, because they addressed one element.

[12] Information about the ComponentSoftware HTMLDiff tool is available at http://www.componentsoftware.com/products/HTMLdiff/.

For each web page, we noted whether an identified problem or an observed modification related to accessibility (i.e., changes to improve access, such as the addition of alt text, ways to skip over links, or initial text in text fields), usability (i.e., changes to improve readability, information seeking, etc., such as adding clearer headings, removing background images, or changing color combinations), or other aspects (i.e., changes not considered as improving usability or accessibility directly, such as adding missing tags, adding metadata, or repairing broken tags). In many cases, a change or problem was relevant for improving both usability and accessibility. For each modified page, we also noted whether the modification was visible or introduced errors (e.g., caused incorrect page rendering or added stray text).

We used descriptive statistics (maximum, median, etc.) to aggregate both the number and the type of problems and modifications for each modified site. We only present results from our site-level analysis in this paper.

Study Results

We examine whether the three tools enabled more comprehensive evaluations or enhanced designers' throughput, based on quantitative measures. We then summarize subjective evaluations of tool effectiveness.

Evaluation Coverage

Our analysis of the number of problems identified in conditions with and without automated tools revealed that, as expected, the automated tools identified more problems than the designers identified, especially when considering the total number of errors identified within each site (Figure 3). (We show in the next section that designers made more design changes when they did not use an automated tool.) The numbers do not show that the tools, excluding Bobby, mostly report the same errors (once for each occurrence). Figure 3 shows that there is not much difference in the number of Bobby-reported problems at the page-level and the number of designer-reported problems, possibly because Bobby summarizes the number of unique problems along with the number of occurrences.

Designers primarily identified higher-level problems that were applicable to the entire site rather than to individual pages, such as "Spell things more clearly...," "...make clearer what the section is describing and where the links will go," and "...multicolored links (bad for vision impairment)...." There were only a few problems identified by both designers and one or more tool, including: add attributes or labels for form controls (Bobby and LIFT tools), reduce the number of links (LIFT-NNG), increase text size (LIFT-NNG), add or remove alt text (Bobby and LIFT tools), and add skip links (LIFT tools). Designers did not enumerate any problems that the Validator identified, but they repaired tags when they modified sites in the manual condition.

Designer Throughput

Although the automated tools identified more problems than the designers identified, study results suggest that they did not enhance the designers' throughput (i.e., their ability to resolve identified design problems within a short amount of time). Figure 4 shows that designers made significantly more changes in the manual condition than in the tool conditions ($p < 0.05$). Designers modified a median of three pages on each site in the manual condition, two pages in both LIFT conditions, one page in the Validator condition, and no pages in the Bobby condition; differences were significant ($p < 0.05$). Analyses of variances (ANOVAs) did not show significant effects due to prior tool usage or design experience.

Surprisingly, designers only modified a total of three pages in the Bobby condition. Participants' comments suggest that they did not think that any modifications were needed, based on the reported problems; this is evident from the relatively higher proportion of designers reporting that they completed site modifications and were satisfied with them in the Bobby condition (Figure 5). It is even more surprising that designers made significantly more modifications in the manual condition than in the LIFT conditions, because the LIFT tool can assist designers with making some repairs, such as changing alt text or image dimensions. There were no significant differences in the number of problems that tools reported before and after modifications.

Figure 6 shows that designers made more holistic changes in the manual condition than in the other tool conditions, except for in the combo condition. Their modifications addressed usability, accessibility, and other problems fairly equally. The figure also shows that designers introduced more errors while making modifications in the manual condition than in the individual tool conditions; designers also introduced errors

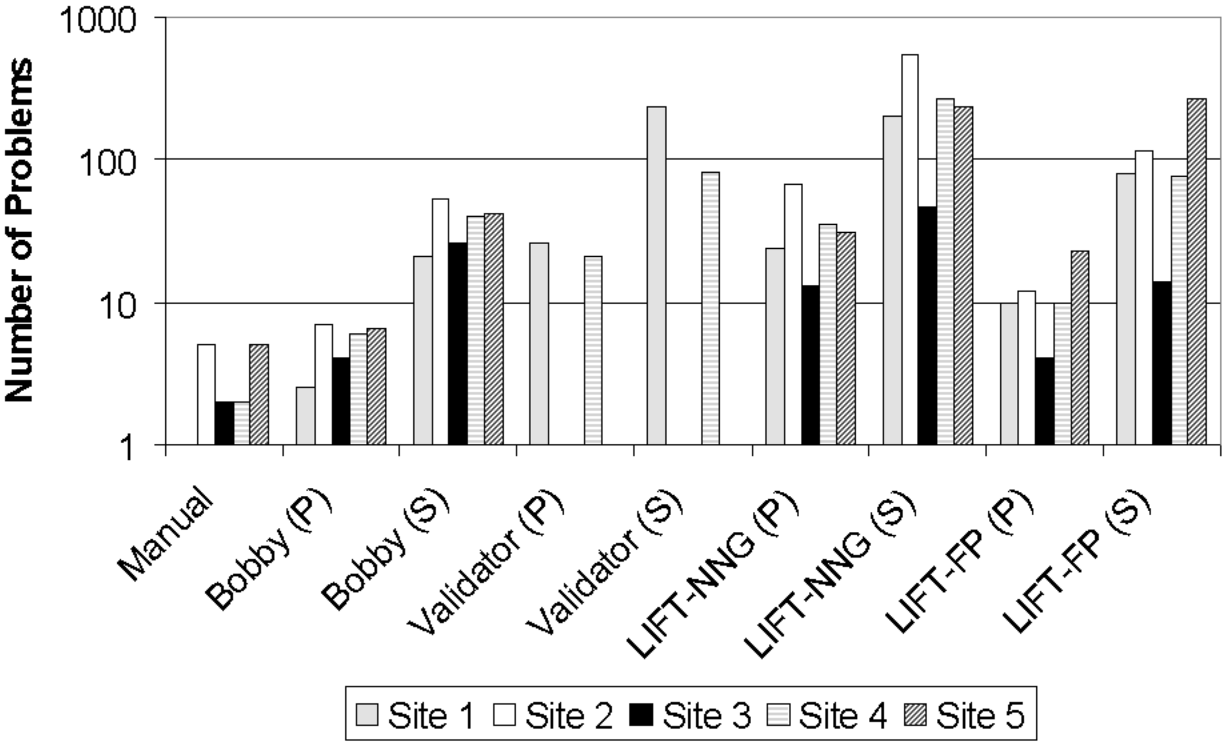

Figure 3: Maximum number of problems identified for each site subsection. For the automated tools, we depict both the median (P) and the total (S) number of problems identified across pages in the site; we report the median due to the small number of pages within each site subsection (Table 1). No designer modified site 1 in the manual condition. Bobby errors were summed together. The Validator failed on sites 2, 3, and 5.

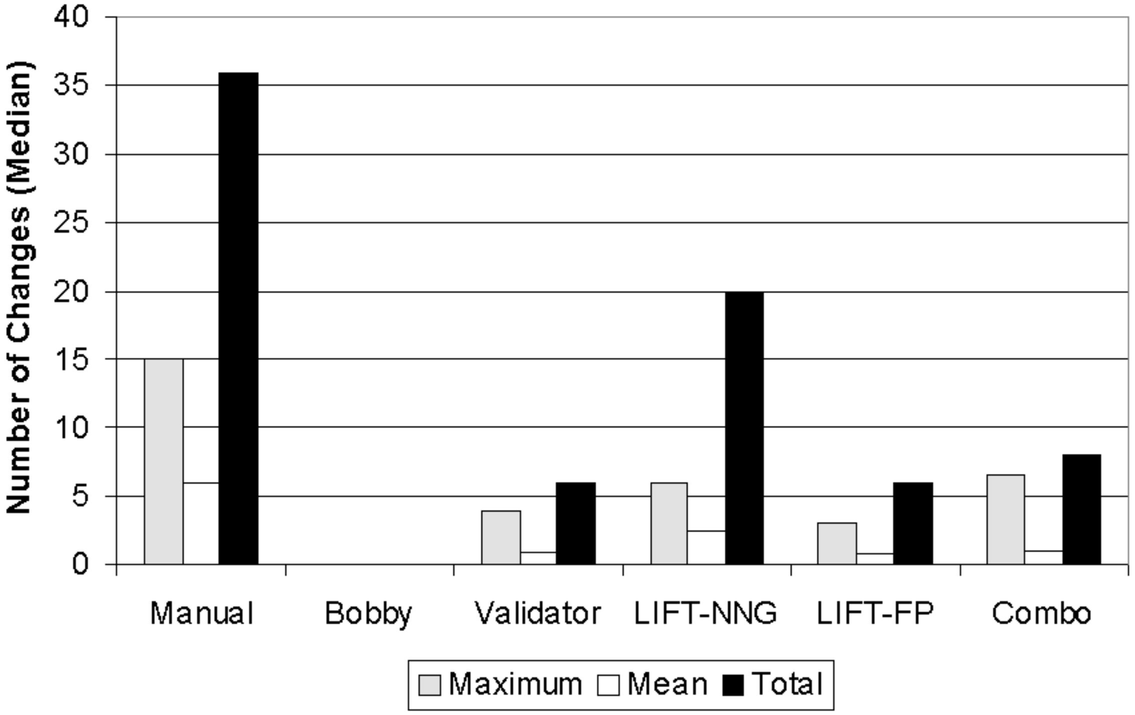

Figure 4: Median number of changes made. For each site, we computed the maximum, mean, and total number of changes made on individual pages. We then computed the median of these site-level measures across sites, which we depict. ANOVAs revealed that all differences are significant ($p < 0.05$).

in the combo condition. Our analysis of the changes made across designers for the same site and the same condition revealed that designers made similar types of changes more often in the manual condition than in the tool conditions. There was typically one page within each site for which multiple designers made one or more similar changes. Results suggest that the tools may or may not result in deterministic changes; designers could interpret reported problems or solutions quite differently.

Tool Effectiveness

We asked designers to rate the universal accessibility of sites both before and after modifications. A related-samples test revealed no significant differences in these ratings, possibly because designers often did not have enough time to complete modifications (Figure 5) or thought that sites needed to be redesigned.

For each tool, designers' responded that they strongly agreed to strongly disagreed with nine statements, including: assessments were accurate and relevant, the tool was easy to use, the tool was extremely helpful in improving universal accessibility, and the tool helped them to learn about effective design practices. There were significant differences in responses to the latter two statements. Responses suggest that Bobby was the most helpful and that the Validator was the least helpful. Similarly, responses suggest that Bobby and LIFT-NNG helped designers to learn about effective practices.

On the debriefing questionnaire, 50% of designers identified Bobby as the most effective tool, and 62.5% identified the Validator as the least effective tool. Surprisingly, 55.6% of designers reported that neither using all three tools nor using one tool was the best usage scenario. Several designers commented about the need to rely on designers' expertise or usability analysis (e.g., "...there are usually embedded corporate reasons for placement of objects...needs to be taken into consideration.").

Web Users' Usage of Modified Sites

The preceding study demonstrated that although the automated evaluation tools identified more errors than designers identified without using them, designers made more design changes in the condition without tools.

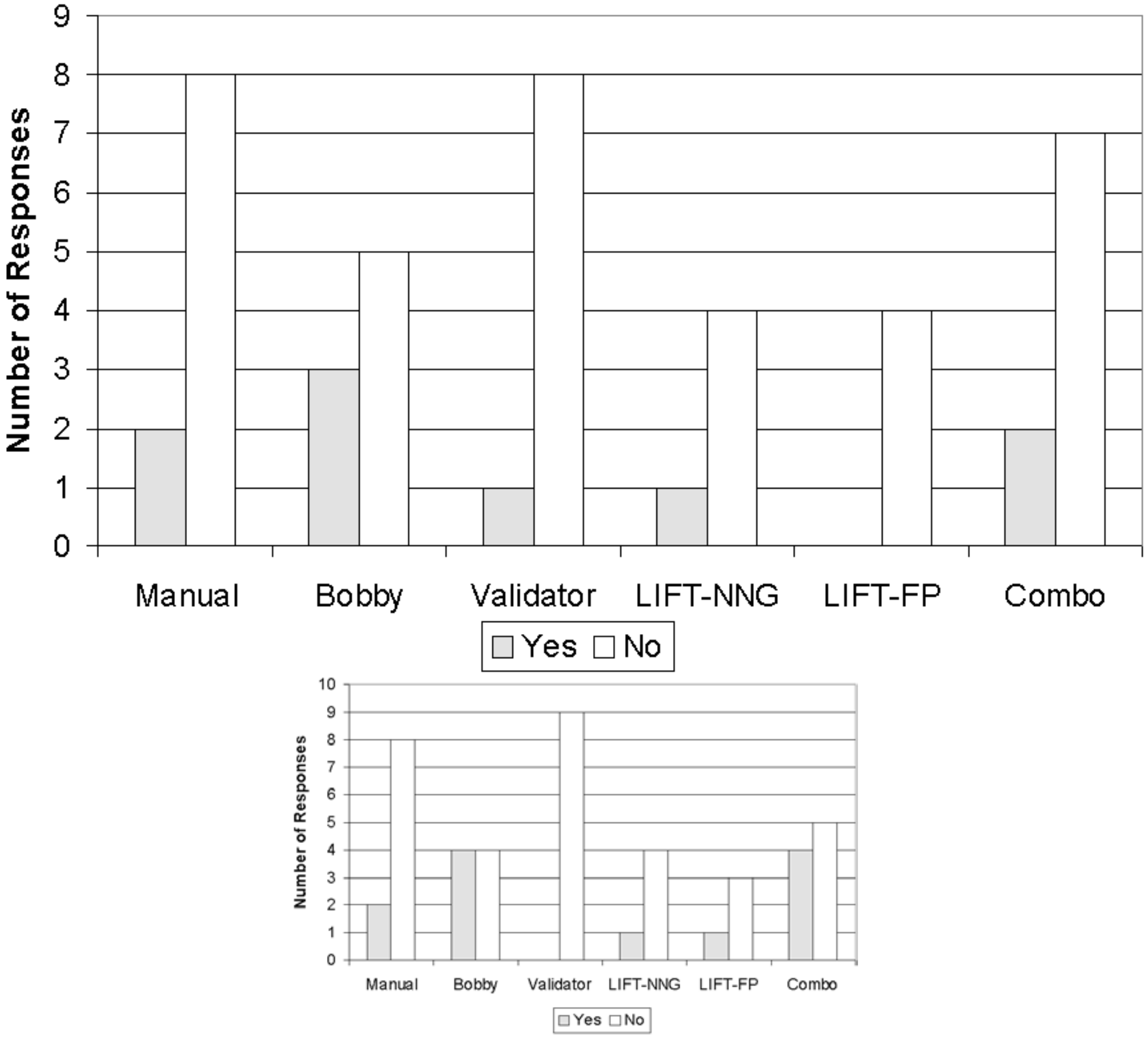

Figure 5: Designers' responses to questions about whether or not they completed their site modifications (top figure) and whether or not they were satisfied with the modifications they made (bottom figure).

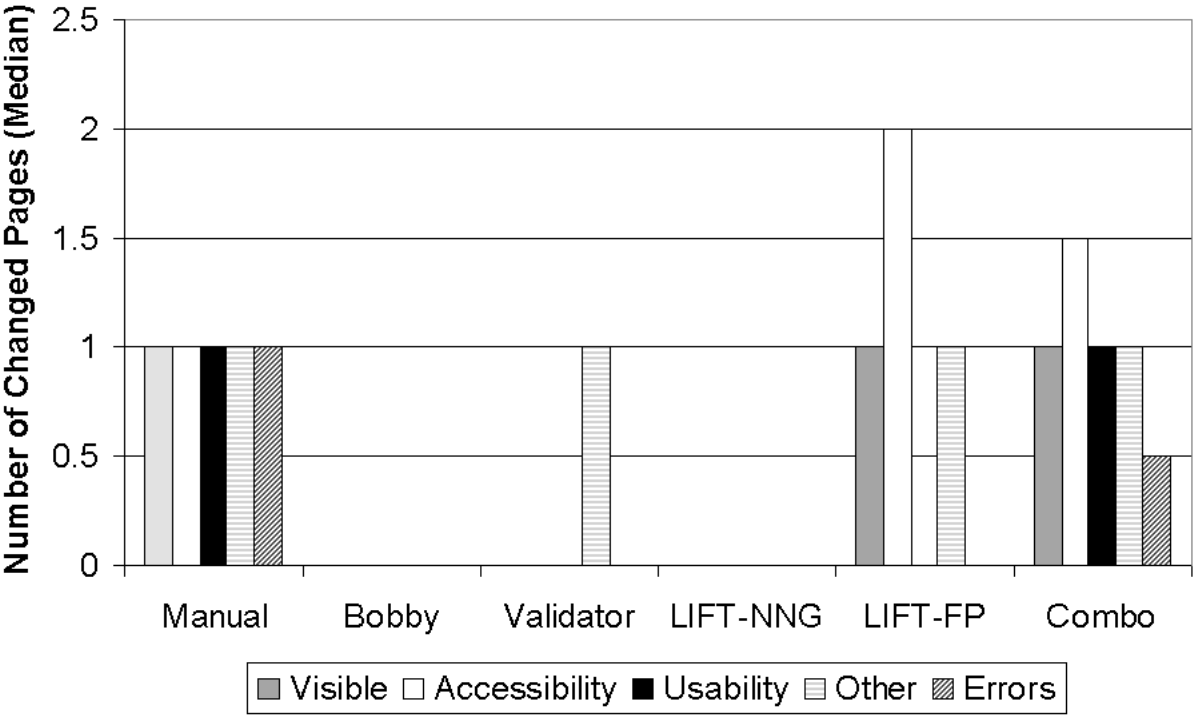

Figure 6: Median number of pages with visible, accessibility, usability, other, and erroneous changes. For each site, we computed the total number of pages with changes of each type. We then computed the median of these site-level measures across sites, which we depict. ANOVAs revealed that the accessibility and error differences are significant ($p < 0.05$).

To determine if the changes resulted in measurably different site experiences, we selected modified versions of each of the five sites for usability testing by users with and without disabilities.

Study Design

We designed a 6x5 between-subjects experiment to determine if there would be any differences in user performance or in subjective ratings for the original and modified versions of the sites (Table 2). We also wanted to understand the types of changes that produced significant differences. We asked users to participate in a ninety-minute or two-hour study session[13] wherein we gave them a scenario for each site and then asked them to browse the site to find specific information (see Table 1). We counterbalanced both the site order and the site version (i.e., modification condition); participants completed tasks on only one version of each site.

Participants noted whether or not they successfully completed each task and evaluated the site's universal accessibility, similarly to the web designers; we also recorded the task completion time for each task. We asked designers to think aloud throughout the study session, and we videotaped sessions and captured screen activity. We observed that participants often thought that they had completed an information-seeking task when they had not; thus, we analyzed server log data to determine and record whether participants actually completed each task. We had disabled browser caching so that server logs would be as accurate as possible.

Participants tested sites modified by all but one of the designers. Table 2 depicts the site versions and the total number of participants who evaluated each version.

Study Participants

There were twenty-two participants in this study. Their ages ranged from 18 to 55, with 63.6% being between the ages of 18 and 35. An equal number of males and females participated. Participants had very

[13] The two-hour sessions were to allow extra time for participants with disabilities to complete the study.

Condition	Site				
	1	2	3	4	5
Original	6 (4)	4 (2)	5 (4)	4 (2)	4 (2)
Manual	–	5 (3)	4 (2)	8 (4)	5 (4)
Bobby	–	–	–	5 (4)	3 (1)
Validator	5 (4)	–	8 (4)	–	5 (3)
LIFT-NNG	7 (3)	5 (3)	–	–	5 (3)
Combo	4 (2)	8 (5)	5 (3)	5 (3)	–

Table 2: Total number of participants who evaluated each site version. The number to the left of each parenthesis denotes the total number of participants, and the number in the parenthesis denotes those participants who had disabilities.

diverse ethnic backgrounds, and the majority were native English speakers, had a college or post-graduate education, were intermediate computer and Internet users, and typically spent ten or more hours online a week. Thirteen participants had a visual (6), hearing (1), moderate mobility (3), learning (1), physical (5), or other (1) disability. Three participants actually had multiple disabilities.

Participants mainly used Netscape Navigator to complete study tasks. Blind participants experienced problems with the JAWS screen reader software and needed to use Internet Explorer instead. Two participants with mobility disabilities used the keyboard exclusively. For participants with visual and physical disabilities, testers read the study tasks aloud and assisted participants with completing the post-task questionnaires. For 89% of the tasks, participants had never used the sites.

Study Results

When we analyzed data across all the sites, ANOVAs showed no significant differences in the participants' task completion success, their task completion times, or their subjective ratings based on the site versions (i.e., modification conditions) as well as on whether participants had disabilities or not. However, our analysis of data for individual sites revealed a few significant differences in results for sites 2, 3, and 5. It is interesting to note that the modifications designers made to these sites in the manual condition improved user performance, yet this was not the case for the modifications designers made in the tool conditions.

Participants spent more time completing tasks on the LIFT-NNG version of site 2 than the original one (median of 286 seconds versus 163). The designer modified all pages in the LIFT-NNG version, but the modifications addressed coding issues (e.g., fixing tags and adding a document type definition) instead of usability and accessibility issues. The designer also introduced an error on the first page, which effected its layout.

Participants rated the manual version of site 3 higher than the original one (3.25 versus 1.4). The designer made extensive visible modifications to four of the six pages in the manual version. Most of the changes addressed usability issues: wrapping long text links in the navigation bar and adding vertical space between them, increasing the font size for links, formatting headings so that they appear more distinctly, adding vertical space between links in a bulleted list, and removing an acronym scheme to reduce potential cognitive overload. The designer did not introduce errors on any pages.

No participant was able to complete the task on the Validator version of site 5, but 80% of participants completed tasks on the manual version. (Only 25% of participants completed tasks on the original version.) The designer only modified one page in the Validator version; changes consisted of adding alternative text to indicate advertisements. There was a blind user in this condition, and the user did not complete the task. The designer modified three pages in the manual version; changes addressed usability and accessibility issues, such as moving relevant links to a prominent position at the top of the page, adding headings to the table of contents for an FAQ, increasing font size, and adding vertical space around links. Unlike the Validator version, the manual version had visible errors on two pages.

Participants only spent a median of 91 seconds completing tasks on the LIFT-NNG version of site 5 versus 186 seconds on the manual version. However, twice as many participants (80%) actually completed the task on the manual version, which explains the extra time. The designer only modified alternative text to indicate spacer images on pages in the LIFT-NNG version. Due to the busy nature of the initial page, participants typically gave up on completing the task. As discussed above, the designer of the manual version made considerable changes to reduce page complexity.

Study Implications

Within the narrow parameters of our study, results suggest that the three automated evaluation tools were not as effective in helping designers to improve web site usability and accessibility as we had hypothesized. However, we consider our results to be preliminary. Perhaps if used less mature sites, studied novice or expert designers, or gave designers more time to learn the tools, to understand the guidelines, and to modify sites, the outcome may be different. Nonetheless, our findings have implications for both researchers and web professionals.

More research needs to be done to validate the guidelines embedded in automated evaluation tools. Fundamentally, results suggest that the guidelines do not examine higher-level issues, such as page complexity or whether text is legible with the color combinations used. The tools should not apply guidelines independently; for example, increasing font size may require a subsequent reduction in the amount of text on a page, but this is not addressed by the tools. The WebTango approach [8] (described briefly in the Background section) provides an example of how to address these issues.

More research needs to be done to make the guidelines and the tools themselves more usable. Guidelines are often too long or too vague for designers who do not have a human factors background to understand. Thus, designers experienced difficulties interpreting them, which in turn interfered with their ability to implement appropriate changes. Chevalier and Ivory's[14] studies on the effects of guideline training sessions provide some guidance on helping designers to apply guidelines. More studies need be conducted to examine the cognitive difficulties web designers encounter in using these tools.

More work needs to be done to improve the presentation of guideline violations, because designers find it overwhelming to wade through long lists of violations, especially when they consist of the same type of violations. Bobby provides a good model for how to minimize overload when presenting this information. Lastly, guidelines need to be context sensitive and perhaps even extensible so that designers can adjust them based on the context in which sites are designed.

Although it should be evident that web professionals cannot rely on the automated evaluation tools alone to improve sites, the results show this to be the case. The studies suggest that the guidelines embedded in the tools, or perhaps the way designers interpret them, may not necessarily improve the usability and accessibility of web sites, at least not more so than when experienced designers rely on their own expertise. Results do suggest that the tools play an important role in educating web professionals about effective design practices.

Conclusions

We examined the effectiveness of the WatchFire Bobby, W3C HTML Validator, and UsableNet LIFT automated evaluation tools. Nine designers used the tools to modify five sites, and twenty-two users (with and without disabilities) completed information-seeking tasks within the original and modified sites. Although the tools helped designers to identify a larger number of potential problems, designers were not effective in interpreting and applying the guidelines. Furthermore, the modifications that designers made based on the tools did not improve user performance or ratings.

Acknowledgements

We wholeheartedly thank the survey respondents and study participants. We also thank the many people who provided assistance with survey announcement and participant recruitment, including Maya Draisin at the International Academy of Digital Arts and Sciences, Jerry Evas of SpiderMetrix.com, Beth Kolko and Mary Deaton of the Technical Communication Department at the University of Washington, Sheryl Burgstahler and the staff of the DO-IT program at the University of Washington, and distribution list moderators (CHI-Web, SIGIA, SIGHCI, and many others). We thank Judy Ramey for allowing us to use

[14] Personal communication, Aline Chevalier and Melody Y. Ivory, *Can novice designers apply usability criteria and recommendations to make web sites easier to use?*, 2002.

the Laboratory for Usability Testing and Evaluation. We thank Chantrelle Nielsen for her invaluable support in facilitating the studies. We also thank Marti Hearst, Jen Mankoff, and the anonymous reviewers.

References

[1] Marilyn Hughes Blackmon, Peter G. Polson, Muneo Kitajima, and Clayton Lewis. Cognitive walkthrough for the web. In *Proceedings of the Conference on Human Factors in Computing Systems*, volume 4 of *CHI Letters*, pages 463–470, Minneapolis, MN, April 2002.

[2] Giorgio Brajnik. Automatic web usability evaluation: Where is the limit? In *Proceedings of the 6th Conference on Human Factors & the Web*, Austin, TX, June 2000.

[3] Helen Brazier and Simon Jennings. Accessible web site design: What the guidelines don't tell you or how not to make a meal of it: The fiction cafe. *LTWorld*, February 5, 1999.

[4] Aline Chevalier and Melody Y. Ivory. Web site designs: Influences of designer's experience and design constraints. Submitted for publication, 2002. Available at http://webtango.berkeley.edu/papers/designers.pdf.

[5] Kara Pernice Coyne and Jakob Nielsen. Beyond alt text: Making the web easy to use for users with disabilities. Nielsen Norman Group Report, 2001.

[6] International Organisation for Standardisation. *ISO9241 Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 11: Guidance on Usability.* International Standard. Geneva, Switzerland, 1998.

[7] Melody Y. Ivory and Marti A. Hearst. State of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516, December 2001.

[8] Melody Y. Ivory and Marti A. Hearst. Statistical profiles of highly-rated web site interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*, volume 4 of *CHI Letters*, pages 367–374, Minneapolis, MN, April 2002.

[9] Gene Lynch, Susan Palmiter, and Chris Tilt. The Max model: A standard web site user model. In *Proceedings of the 5th Conference on Human Factors & the Web*, Gaithersburg, Maryland, June 1999.

[10] C. Stephanidis, D. Akoumianakis, M. Sfyrakis, and A. Paramythis. Universal accessibility in HCI: Process-oriented design guidelines and tool requirements. In *Proceedings of the 4th ERCIM Workshop on User Interfaces for All*, Stockholm, Sweden, October 1998.